SUPPLEMENTARY INFORMATION
for

Codon influence on large-scale protein expression correlates with *E. coli* mRNA levels

Grégory Boël [1,2,3], Reka Letso [1,2†], Helen Neely [1,2†], W. Nicholson Price [1,2†¶]
Kam-Ho Wong [1,2], Min Su [1,2], Jon Luff [1,2], Mayank Valecha [1,2],
John K. Everett [2,4], Thomas B. Acton [2,4], Rong Xiao [2,4],
Gaetano T. Montelione [2,4,5], Daniel P. Aalberts [6§], and John F. Hunt [1,2§]

[1] Department of Biological Sciences, 702A Fairchild Center, MC2434, Columbia University, New York, NY 10027, USA;

[2] Northeast Structural Genomics Consortium;

[3] CNRS FRE3630, Institut de Biologie Physico-Chimique, 13-rue Pierre et Marie Curie, 75005 Paris, France;

[4] Department of Molecular Biology and Biochemistry, Center for Advanced Biotechnology and Medicine, Rutgers, the State University of New Jersey, Piscataway, NJ 08854, USA;

[5] Department of Biochemistry, Robert Wood Johnson Medical School, Rutgers, the State University of New Jersey, Piscataway, NJ 08854, USA;

and

[6] Department of Physics, Williams College, Williamstown, MA 01267, USA.

[†] These authors contributed equally to the work reported in this paper.

[¶] Present addresses: WNP, University of New Hampshire School of Law, 2 White Street, Concord, NH 03301, USA.

[§] To whom correspondence may be addressed:
e-mail jfh21@.columbia.edu, voice (212)-854-5443, FAX (212)-865-8246 (JFH);
e-mail aalberts@williams.edu, voice (413)-597-3520, FAX (413)-597-4116 (DPA).

Running Title: Codon influence tracks mRNA level in *E. coli*.

SUPPLEMENTARY INTRODUCTION

While it is clear that variations in mRNA sequence play an important role in regulating protein expression in organisms from *E. coli* to humans, many different mechanistic hypotheses have been proposed to explain these effects[1,2], and their influence on translation efficiency remains unclear and in some cases controversial. Studies on several proteins[3,4] and later genomic studies[5-8] concluded that stable mRNA folding in the 5' region (head) of a gene, but not downstream in the coding sequence, can attenuate translation in *E. coli*, presumably due to inhibition of the assembly of the 70S ribosomal initiation complex[3,4]. However, no prior study has systematically quantified the influence of mRNA folding energy at all locations in a gene on protein-translation efficiency. Furthermore, substantial uncertainty exists concerning the influence of synonymous codons on translation efficiency[5-7,9-17], the mechanistic basis of such effects, and their relationship to mRNA folding effects[3-8]. Ribosome-profiling studies concluded that the net translation-elongation rate is effectively constant for *E. coli* mRNAs, irrespective of codon usage[10,15,18]. This finding challenges the assumption that differences in the translation rate of synonymous codons influence protein expression, an assumption underlying much of the codon-usage literature[1,2,8,19-21], but no alternative mechanism has been proposed to explain the many experiments showing that changes in codon usage can produce dramatic alterations in protein expression[2,11,22].

Uncertainty furthermore exists concerning which codon-related properties are beneficial *vs.* detrimental for protein expression[5-7,9-12,14-17]. For example, more homogeneous codon usage has been proposed alternatively to enhance[9] or reduce[23] translation efficiency. Much of the codon-usage literature focuses on inefficient translation of a set of rare codons[24] in the *E. coli* genome[21,24,25], especially the AUA codon for ile[12,26] and the AGA, AGG, and CGG codons for arg[11,22,27,28]. On this basis, it is widely assumed that genomic codon-usage frequency, which parallels tRNA pool level[25,29], influences translation efficiency and that infrequent codons are translated inefficiently[11,12]. However, the expression of β-galactosidase[30,31] and a fluorescent reporter protein[5] are increased when the head of the gene contains the rare codons most cited as a barrier to translation. This effect was interpreted to reflect tolerance for inefficient codon usage in the head to prevent stable mRNA folding that would attenuate translation[5]. Although some computational[5,7] and experimental[31] studies support this interpretation, there has been limited mechanistic characterization of the influence or interplay of these parameters, and alternative theories[16,32,33] suggest that, under some physiological conditions, rare codons can accelerate ribosomal decoding kinetics and thereby enhance translation efficiency via mechanisms unrelated to mRNA folding.

The evolutionary biology literature focuses on a different correlate of genomic codon-usage frequency, which is accuracy in protein synthesis[14,34,35]. Biochemical studies suggest that more frequent codons should be translated more accurately because the levels of their cognate tRNAs are systematically higher, and competition from near-cognate tRNAs is the major cause of translational errors[25,29,36]. Usage of more frequent codons is enhanced at more conserved sites in proteins[17,35], presumably because more accurate translation[37] at such sites promotes greater

evolutionary fitness[14,38]. While lower frequency codons also could be translated less efficiently[19], a systematic correlation between these parameters has yet to be demonstrated.

SUPPLEMENTARY RESULTS

Additional information on generation of the large-scale protein-expression dataset. Proteins were transcribed from the bacteriophage T7 promoter in pET21, a 5.4 kb pBR322-derived plasmid harboring an ampicillin resistance marker[39]. We used a bacteriophage polymerase to drive transcription to minimize potentially confounding effects from the coupling of translation to transcription by the native *E. coli* RNA polymerase[40,41]. Protein expression[39] was induced overnight in chemically defined medium at 18 ˚C in *E. coli* strain BL21(DE3), which contains a single IPTG-inducible gene for T7 polymerase. This strain also contained pMGK, a 5.4 kb pACYC177-derived plasmid that harbors a kanamycin resistant gene, a single copy of the *lacI* gene, and a single copy of the *argU* gene encoding the tRNA cognate to the rare AGA codon for arg. All proteins were expressed with the same eight-residue C-terminal affinity tag (with sequence LEHHHHH) appended after the final amino acid in the native protein sequence. Technical details of the large-scale expression experiments are described in the *Methods*.

Characteristics of highly expressed genes in the protein-expression dataset. Protein expression tends to decrease when mRNA A/U/G/C content becomes non-uniform (**Extended Data Fig. 2b-e**). The reading-frame dependency of these base-composition effects suggests that codon properties influence expression, an inference supported by strong correlations between the frequencies of some codons and protein expression level. The GAA codon for Glu shows the strongest positive correlation, whereas the frequency of the synonymous GAG codon is equivalent for all expression scores (**Fig. 1a,b,e**). The AUA codon for Ile shows one of the strongest negative correlations, whereas the synonymous AUU codon shows a weakly positive correlation (**Fig. 1c-e**). These trends naïvely suggest differences in translation efficiency, but the multivariate statistical analyses below indicate their origin is more complex. Nonetheless, some of these effects likely derive from differences in codon translation efficiency, including the strong expression-attenuating effect of adjacent AUA codon pairs (**Extended Data Fig. 2f**). In contrast, expression level is not influenced by the Codon Adaptation Index[21] (CAI), except for a modest decrease at the lowest ~10% of CAI values in our dataset (**Extended Data Fig. 2g**), and neither the tRNA Adaptation Index[33] (**Extended Data Fig. 2h**) nor the frequency of AGGA[15] (**Extended Data Fig. 2i-j**), which matches the core of the Shine-Dalgarno ribosome-binding sequence, are significantly correlated with expression.

The distributions of predicted mRNA folding energies[42] show systematic differences between proteins with different expression scores. Expression is attenuated by increasingly stable folding (*i.e.*, decreasing partition-function free energy of folding[42]) in the first 48 nucleotides in the coding sequence[3,4], which we refer to as the head of the gene (**Fig. 1h**). Our data calibrate this previously observed effect[3,4,6-8] , showing an ~1/*e* reduction in the odds of high expression when folding energy in the head (ΔG_H) reaches -15 kcal/mol. The strength of this correlation is increased modestly by including the 5'-UTR or untranslated region (ΔG_{UH} in **Fig.**

1f,h). Unexpectedly, $<\Delta G_T>$, the average value of the predicted folding energy in the tail of the gene (*i.e.*, nucleotide 49 through the last nucleotide preceding the C-terminal tag) shows a non-linear influence, with both high and low values systematically attenuating expression (**Fig. 1g,h**). Although these observations suggest that excessively stable or unstable folding in the tail attenuates expression, analyses below indicate these effects also have more complex origins.

Several global sequence parameters systematically vary with expression. Very long or short proteins show lower expression (**Fig. 1i-j**). Increasing codon (**Extended Data Fig. 2m**) or amino acid (**Extended Data Fig. 2k-l**) repetition rate (average frequency of recurrence) correlates with lower expression, as does higher statistical entropy in codon (**Extended Data Fig. 2o**) or amino acid (**Extended Data Fig. 2n**) sequence. The amino acid repetition rate *r* is the most influential (**Extended Data Table 1a**) of these mutually correlated parameters, suggesting local repetition of the same amino acid may attenuate expression.

Experimental dissection of the mRNA features controlling high-level protein expression. To investigate whether codon usage in the tail can influence protein expression, we retained the native head sequences and optimized the codons exclusively in the tails of four genes using the 6AA method ($WT_H/6AA_T$ in **Fig. 5b** and **Extended Data Fig. 6b**). Tail optimization increases expression of all four of these target proteins, although the extent of improvement varies substantially. For SRU_1983, protein expression increased only slightly. However, the other target proteins showed either significant (SCO1897) or very large (RSP_2139 and APE_0230.1) increases in expression normalized to total cell protein, verifying the inference from our computational analyses that codon content in the tail can have a powerful influence on protein-expression level.

We also tested the relative influence of codon usage *vs.* mRNA folding in the head by constructing genes with identical tails but different heads that were codon-optimized using the 31C method while either optimizing (31C-FO_H with maximized ΔG_{UH}) or deoptimizing (31C-FD_H with minimized ΔG_{UH}) their calculated free energies of folding (31C-$6AA_T$ constructs in **Fig. 5b** and **Extended Data Fig. 6b**. The 31C-FO heads greatly improved expression of three of four proteins evaluated – RSP_2139, SRU_1983, and SCO1897. The 31C-FO heads for these proteins combined with either native or 6AA-optimized tails produce similarly high levels of expression (**Fig. 5b**). The only protein that did not show a significant improvement upon head optimization, APE_0230.1, is already expressed at an exceedingly high level when the tail is optimized while retaining the native head sequence.

Deoptimizing head folding yielded different results for the four target proteins that paralleled their calculated free energies (**Fig. 5b** and **Extended Data Fig. 6b**). There were large differences between these proteins in the lowest ΔG_{UH} that could be achieved in synonymous heads constructed using the A/U-rich 31C codon set, providing another example of coupling between codon usage and more global physicochemical properties of mRNA sequences. The most stably folded 31C-FD head (RSP_2139 with ΔG_{UH} = -47 kcal/mol) eliminates the very high expression produced by the synonymous 31C-FO head (ΔG_{UH} = -37 kcal/mol), verifying the conclusion from our modeling studies (**Fig. 4** and **Extended Data Table 1a**) and prior literature that stable head folding can block protein expression. The 31C-FD head for SRU_1983

(ΔG_{UH} = -41 kcal/mol) also decreases expression compared to the synonymous 31C-FO head (ΔG_{UH} = -34 kcal/mol). In contrast, the comparatively less stably folded 31C-FD heads for SCO1897 (ΔG_{UH} = -38 kcal/mol) and APE_0230.1 (ΔG_{UH} = -32 kcal/mol) produced equivalent expression to the synonymous 31C-FO heads (ΔG_{UH} = -29 kcal/mol and -27 kcal/mol for SCO1897 and APE_0230.1, respectively). However, the codon-optimized 31C-FD heads for SRU_1983 (ΔG_{UH} = -41 kcal/mol) and SCO1897 (ΔG_{UH} = -38 kcal/mol) both increased expression compared to less folded native heads (ΔG_{UH} = -34 kcal/mol and -35 kcal/mol for SRU_1983 and SCO1897, respectively), supporting our computational inference (**Fig. 4, Extended Data Fig. 5,** and **Extended Data Table 1a**) that codon content in the head can strongly influence protein expression.

SUPPLEMENTARY DISCUSSION

Use of heterologous genes to study mRNA sequence features controlling protein expression. The predominance of heterologous genes in our large-scale dataset has several advantages for interrogating the biochemical and physiological mechanisms controlling protein expression in *E. coli*. First, experiments on heterologous genes should reduce or eliminate effects from gene/protein-specific regulatory systems in *E. coli*. Second, taking genes from a variety of heterologous sources provides much greater diversity in sampling of codon-space than genes from *E. coli* or any other single organism, and, in practice, our dataset provided greater codon diversity than achieved in previous studies using synthetic genes to examine codon-usage effects[5,6,15]. Finally, challenging the biochemical apparatus in a given organism with sequences that have not evolved under selective pressure in that organism should reduce the influence of parallel selective effects acting on sequential steps in a physiological pathway. Parallel selection can create evolutionary surrogate effects in the form of significant sequence correlations that do not reflect a direct mechanistic effect. Regulation of protein expression minimally involves the interplay of transcription, translation, RNA degradation, and protein degradation. Endogenous *E. coli* genes are likely to have sequence features influencing some of these interconnected processes but not others, which can produce surrogate effects in analyses evaluating correlations between gene sequences and experimental outcomes.

For example, sequence features enhancing transcription could be enriched in genes that are translated efficiently, because both efficient transcription and translation are needed to achieve high protein expression. As explained in the main text, we used bacteriophage T7 RNA polymerase rather than the endogenous *E. coli* RNA polymerase to drive transcription to minimize the influence of transcriptional mechanisms on protein expression level. However, a dataset comprising exclusively genes from *E. coli* or closely related organism could show statistical correlations between sequence features controlling transcription and protein expression level, even if those features do not directly influence expression in our experiments, because of the co-evolution of features controlling these sequential steps in the physiological protein-expression pathway. To the extent that there is divergence in the biochemical and physiological properties of the source organisms, evaluating the expression of heterologous genes in *E. coli* will reduce the influence of indirect sequence correlations and evolutionary surrogate effects of

this kind. Therefore, only biochemical effects that are universally conserved among the diverse source organisms should produce strong surrogate effects in our experiments due to parallel selection for sequence features influencing sequential steps in the expression pathway.

As discussed further below, disentangling sequence features influencing translation efficiency and mRNA stability is substantially more difficult than disentangling those influencing transcription, because both translation efficiency and mRNA stability influence protein expression level in *E. coli* even when transcriptional dynamics are identical. Evaluating expression of proteins from heterologous proteins is more likely to reveal effects related to translation than mRNA stability, because nucleotide sequences evolve significantly more rapidly than protein sequences, and the translation machinery is believed to be more strongly conserved evolutionarily than mRNA decay systems. Nonetheless, some of the sequence features controlling mRNA stability could be broadly conserved evolutionarily, and universally conserved biochemical mechanisms will influence statistical analyses performed on any dataset examining net protein expression level from naturally evolved genes, irrespective of the source of the gene sequences. Therefore, mechanistically resolved *in vitro* follow-up experiments are required to verify any biochemical hypotheses derived from large-scale datamining studies.

In the current paper, we performed such *in vitro* follow-up experiments on a set of synthetic genes designed using an algorithm derived from our computational analyses of the large-scale dataset that we generated using primarily heterologous genes to interrogate the biochemical and physiological mechanisms controlling protein expression in *E. coli*. We designed genes for four heterologous proteins without *E. coli* orthologs to avoid effects from gene/protein-specific regulatory systems in *E. coli*. We also designed a gene for the *E. coli* protein YacQ to verify that an endogenous protein gives consistent results. All of these synthetic genes exhibit similar behavior in all of our biochemical assays, providing important support for the validity of our hypotheses concerning mechanisms underlying the control of protein expression in *E. coli*.

Additional support for the success of our experimental strategy employing primarily heterologous genes to interrogate the mRNA sequence features controlling protein expression in *E. coli* comes from comparing the performance of our computational model on the 95 *E. coli* genes *vs.* the 6,253 heterologous genes in our large-scale dataset. Our multiparameter generalized linear logistical regression model predicting high *vs.* no expression performs similarly on both sets of genes (data not shown). The performance of the model is significantly better for the *E. coli* genes than the heterologous genes in the top expression category ($p = 0.004$ for a 2-sided T-test for the distribution of predicted probabilities), even though they represent only ~1.5% of the dataset. This result is consistent with our experimental approach being an effective way to capture fundamental features of *E. coli* biochemistry and physiology.

Validation of inferences from computational modeling. The foundation of the work reported in this paper is simultaneous multi-parameter computational modeling, which is a powerful tool because, in principle, it can resolve the relative influence of cross-correlated parameters (*e.g.*, many pairs of parameters evaluated in **Extended Data Fig. 3** including codon content and predicted RNA-folding energy[42]). However, there can be noise in these estimates, and, as

discussed in the preceding section, the apparent influence of some parameters can reflect evolutionary surrogate effects, *i.e.*, the "hidden" influence of cross-correlated sequence parameters not included in the analysis. For example, if evolution constrains more highly expressed proteins to be more soluble, there could be a positive correlation between protein-expression level and the frequency of codons for solubility-enhancing amino acids, even if these amino acids do not increase protein-translation efficiency. These considerations reinforce the need discussed above to validate computational inferences using experiments that rigorously evaluate specific biochemical mechanisms. Our *in vitro* translation experiments comparing native and computationally designed genes (**Fig. 5c and Extended Data Fig. 6c**) verify that the most influential mRNA sequence features identified in our multi-parameter computational model (**Model M** in **Extended Data Table 1a** and **Figs. 3-4**) directly modulate translation, ruling out substantial interference from statistical noise, hidden variables, surrogate effects, or other latent systematic errors.

Although the experimental data presented in this paper strongly support the major conclusions from our computational modeling studies, the details of these studies require further validation, both to ensure their quantitative accuracy and to elucidate the underlying molecular mechanisms. A high priority in this area will be to evaluate whether our new codon-influence metric (colored symbols in **Fig. 3a**) accurately describes the relative translation efficiencies of the different amino acids and the synonymous codons for the same amino acid. The broad features of this metric are validated by the very strong correlations of its average value with global physiological protein and mRNA levels *in vivo* in *E. coli* (**Fig. 6**), but the differences in the values for some synonymous codon pairs (**Fig. 3a**) are not themselves statistically significant within our computational model (**Model M** in **Extended Data Table 1a**). Protein expression experiments *in vivo* and high-resolution enzymological studies of protein synthesis *in vitro*[43-46] will be needed to critically evaluate the quantitative details of our new codon metric and to elucidate more clearly its mechanistic origin.

Disentanglement of parallel selective effects operating on different biochemical process. As discussed in the two previous sections, mechanistically resolved *in vitro* experimentation is essential to verify hypotheses attributing sequence correlations to a specific biochemical effect, because indirect evolutionary couplings and parallel selection operating on sequential steps in a pathway can create surrogate effects, *i.e.*, significant sequence correlations in naturally evolved genes that do not reflect a direct mechanistic effect. The failure to appreciate the importance of such effects pervades most prior literature on codon effects and likely accounts for significant biases and misinterpretations. These considerations highlight the importance of the *in vitro* transcription and translation assays using purified components that are presented in this paper, because these assays represent the most rigorous approach to verifying that the strong codon effects identified our in statistical analyses have a mechanistic effect on protein translation efficiency. In contrast, most of the codon-efficiency metrics used in prior literature on this topic were not validated in biochemical experiments of this kind, meaning that they could potentially reflect in part indirect correlations.

The well-known Codon Adaptation Index (CAI) is one example of such a metric. The logic behind this metric is that the most highly expressed proteins should have efficiently

translated codons. However, such proteins also need to be encoded by transcripts that are efficiently synthesized and resistant to mRNA degradation, so a metric like the CAI is likely to capture an unpredictable mixture of sequence features influencing all of the sequential process required to achieve high protein expression. Notably, we observe that our new codon influence metric fully captures the influence of the *E. coli* CAI plus 7-fold additional deviance in our large-scale protein expression dataset (**Extended Data Table 1**), but CAI captures about twice as much deviance as our new codon metric in preliminary modeling of the *E. coli* mRNA lifetime dataset presented in a recently published global profiling study[47] (data not shown). These observations suggest that the CAI may capture more directly mechanistic effects related to mRNA decay than translation. *In vitro* experimentation of the kind presented in this paper is essential to bridge the gap between any observed evolutionary sequence correlations and reliable inferences concerning biochemical mechanism. An extensive amount of additional *in vitro* experimentation will be required to follow-up on the results presented in this paper and sort out the nature and interdependencies of sequence features influencing transcription, translation, and mRNA degradation.

Two recently published papers suggest that variations in synonymous codon usage influence mRNA decay rate in yeast[48,49], echoing an important conclusion of the studies presented in this paper on *E. coli*. One of these papers[49] reports a correlation in *Saccharomyces cerevisiae* between mRNA lifetimes and a theoretical metric called the tRNA Adaptation Index[33,50] (tAI), and it also reports single-parameter correlations between mRNA lifetimes and the frequencies of some codons. We demonstrate in **Fig. 3a** and **Extended Data Fig. 4b-c** in this paper that single-variable analyses of this kind on our dataset yield misleading conclusions concerning the effects of many codons, because they are strongly biased by evolutionary surrogate effects caused by cross-correlations in the codon content of the genes (**Extended Data Fig. 3**). We also demonstrate in our manuscript that the tAI for *E. coli* is minimally correlated with protein expression in our large-scale dataset (**Extended Data Figs. 2h** and **4g** and **Extended Data Table 1b**). While translational dynamics could have significant differences in yeast compared to *E. coli*, the tAI has not been demonstrated to influence translation efficiency in an *in vitro* system. Given the complex interplay of mechanistic factors influencing sequence evolution, as discussed above, it will be import in the future to verify that the sequence features that correlate with changes in mRNA lifetime in yeast have parallel effects on translation efficiency assayed using *in vitro* translation reactions that decouple this process from other biochemical processes influencing protein expression level and mRNA stability *in vivo*.

Possible mechanisms coupling codon content to mRNA level and lifetime. Several molecular mechanisms could explain the observed correlations between codon content and *in vivo* mRNA concentration and lifetime in *E. coli* (**Fig. 6**). We hypothesize they reflect a kinetic competition between protein elongation and mRNA degradation that is modulated by ribosomal elongation dynamics (*i.e.*, the sequential binding and conformational processes involved in amino-acyl-tRNA selection, peptide-bond synthesis, and tRNA/mRNA translocation). The bacteriophage T7 RNA polymerase used in our experiments synthesizes mRNA too rapidly for translating ribosomes to keep up, making the resulting transcripts insensitive to transcription-translation[40,41]. Therefore, our observation that inefficiently translated mRNAs produced by T7 polymerase are fragmented and have lower concentrations *in vivo* (**Fig. 5d**) is likely to reflect enhanced

degradation. This reasoning, combined with the global correlations between s_{All} and mRNA levels (**Fig. 6a-b**) and lifetimes (**Fig. 6c-d**) in *E. coli* and the tendency of expression-attenuating codons to eliminate protein expression entirely in our large-scale dataset (**Fig. 1a-e**), suggests that mRNA degradation is controlled in part by ribosomal elongation dynamic[51-55]. Several biochemical systems mediate recycling of ribosomes stalled due to protein synthesis/folding problems[56] or mRNA truncation[57]. In eukaryotes, this "No-Go" decay pathway involves the Dom34, Hbs1[58], and ABCE1[59] proteins, whereas in *E. coli*, similar activities are mediated by unrelated systems including the tmRNA pathway[27,56,57], ArfA, YaeJ[60], and RF3[27]. These prokaryotic mRNA quality-control systems are candidates to participate in the mRNA decay process that we hypothesize to be coupled to codon-dependent variations in ribosomal elongation dynamics. Recent reports suggest a similar coupling exists in yeast[48,49].

A coherent model for the biochemical influence of synonymous codon variations. Based on the data in this paper, we hypothesize that inefficiently translated codons attenuate protein expression in two distinct but interrelated ways, first by reducing translation efficiency and thus the yield of protein from an mRNA molecule, and second by enhancing the rate of degradation of that mRNA molecule[40,51-53,61-63]. Inefficiently translated codons could also promote premature termination of mRNAs synthesized by *E. coli* RNA polymerase[64,65], which would also lead to a reduction in steady-state concentration. Overall, the steady-state level of each mRNA is controlled by a dynamic balance between its transcription-initiation rate, which should not depend directly on codon usage, and its premature transcription-termination and decay rates. We hypothesize that a significant influence of codon usage on the rates of the latter two processes creates the strong correlation that we observe between codon content and physiological mRNA levels and lifetimes in E. coli (**Fig. 6**). This proposed feedback between translation efficiency and mRNA level will amplify the influence of codon usage and perhaps also other translational regulatory phenomena on protein expression level, creating a physiologically important but heretofore under-appreciated linkage between translation efficiency and mRNA metabolism.

It is worthwhile to compare this model to the results obtained in recent *in vivo* ribosome-profiling experiments. Such experiments conducted in *E. coli* have raised significant questions about the influence of codon usage on protein expression, because they have shown homogeneous occupancy of the mRNA within each open reading frame (ORF) and a strong correlation between the level of ribosome-occupied ORFs and the concentrations of the encoded proteins[10,15]. These observations imply that ribosomes elongate proteins at a similar rate on most mRNA templates, irrespective of codon usage. However, changes in synonymous codon usage can clearly modulate protein expression level *in vivo*[1,2,5,8,11,18-20,22,27,33,51,52,66-72], and this phenomenon has been attributed in prior literature to codon-dependent variations in mRNA translation rate by ribosomes[11,27,56,67,69,73,74]. This apparent inconsistency between contemporary genome-scale experimentation and a large body of prior literature in molecular biology remains unresolved. The mechanistic model presented above helps to resolve this conundrum, because the inferred influence of codon usage on steady-state mRNA level can lead to a reduction in protein expression from an mRNA molecule irrespective of its translation-elongation rate. As long as most ORFs are translated many times before experiencing an internal codon-dependent event that leads to very rapid processive mRNA degradation, this mechanism is consistent with relatively homogeneous ribosome occupancy within each ORF, as observed in the ribosome-

profiling experiments[10,15]. The level of each ribosome-occupied ORF captures the combined influence of the translation-initiation rate and the steady-state concentration of the corresponding mRNA, which is likely to explain the close correspondence between the concentration of each protein and the level of the ribosome-occupied ORF[10,15]. Codon-dependent variations in mRNA degradation rate will reduce steady-state mRNA concentration, and this effect should be captured by existing ribosome-profiling data even though they do not show widespread codon-dependent variations in translation efficiency[3,18].

On the other hand, several considerations raise questions about the accuracy of this conclusion and the related ribosome-profiling data, including the results from our *in vitro* translation experiments on mRNAs from native *vs.* computationally optimized genes (**Fig. 5c and Extended Data Fig. 6c**). These results reinforce a wide variety of results in previous literature[1,5,11,12,40,51] that led to the general assumption that there are significant codon-dependent variations in translation-elongation efficiency. Given the complexities of the biochemical and evolutionary processes that influence mRNA translation that are outlined above, carefully controlled experiments *in vivo* and *in vitro* will be required to achieve a reliable understanding of codon-dependent variations in translation efficiency and their relationship to mRNA stability. It was widely assumed in prior literature that codon-dependent variations in translation efficiency are attributable to slower accommodation on the ribosome of tRNAs present at lower concentrations in the cell[11,12,19,25,29], which causes slower execution of the translation-elongation cycle for the corresponding codons. The lack of a significant correlation between our new codon-influence metric and tRNA pool levels (**Extended Data Fig. 4h**) raises questions concerning this mechanistic model and suggests that the stereochemical features and allosteric consequences of codon-tRNA interaction are likely to make important contributions to codon-dependent variations in translation efficiency. Future research will be needed to elucidate these effects and also to establish whether codon-related variations in mRNA level (**Fig. 6a-b**) are mediated by altered protection of mRNAs by translating ribosomes or instead by direct recruitment of RNases to ribosomes[75] interacting with inefficiently translated codons or perhaps even by activation of an intrinsic RNase activity in the ribosome itself[76]. Therefore, the results reported in this paper highlight new problems to be investigated in addition to providing new insights and new tools for such studies that lie near the core of the central dogma of molecular biology.

SUPPLEMENTARY REFERENCES

1 Spencer, P. S., Siller, E., Anderson, J. F. & Barral, J. M. Silent substitutions predictably alter translation elongation rates and protein folding efficiencies. *J Mol Biol* **422**, 328-335, (2012).

2 Gingold, H. & Pilpel, Y. Determinants of translation efficiency and accuracy. *Mol Syst Biol* **7**, 481, (2011).

3 Kozak, M. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene* **361**, 13-37, (2005).

4 Shakin-Eshleman, S. H. & Liebhaber, S. A. Influence of duplexes 3' to the mRNA initiation codon on the efficiency of monosome formation. *Biochemistry* **27**, 3975-3982, (1988).

5 Goodman, D. B., Church, G. M. & Kosuri, S. Causes and Effects of N-Terminal Codon Bias in Bacterial Genes. *Science*, (2013).

6 Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-sequence determinants of gene expression in Escherichia coli. *Science* **324**, 255-258, (2009).

7 Bentele, K., Saffert, P., Rauscher, R., Ignatova, Z. & Bluthgen, N. Efficient translation initiation dictates codon usage at gene start. *Mol Syst Biol* **9**, 675, (2013).

8 Tuller, T., Waldman, Y. Y., Kupiec, M. & Ruppin, E. Translation efficiency is determined by both codon bias and folding energy. *Proc Natl Acad Sci U S A* **107**, 3645-3650, (2010).

9 Cannarozzi, G. *et al.* A role for codon order in translation dynamics. *Cell* **141**, 355-367, (2010).

10 Li, G. W., Burkhardt, D., Gross, C. & Weissman, J. S. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell* **157**, 624-635, (2014).

11 Chen, G. T. & Inouye, M. Role of the AGA/AGG codons, the rarest codons in global gene expression in Escherichia coli. *Genes Dev* **8**, 2641-2652, (1994).

12 Caskey, C. T., Beaudet, A. & Nirenberg, M. RNA codons and protein synthesis. 15. Dissimilar responses of mammalian and bacterial transfer RNA fractions to messenger RNA codons. *J Mol Biol* **37**, 99-118, (1968).

13 Price, W. N. *et al.* Large-scale experimental studies show unexpected amino acid effects on protein expression and solubility in vivo in E. coli. *Microbial Informatics and Experimentation* **1**, 6, (2011).

14 Wallace, E. W., Airoldi, E. M. & Drummond, D. A. Estimating selection on synonymous codon usage from noisy experimental data. *Mol Biol Evol* **30**, 1438-1453, (2013).

15 Li, G.-W., Oh, E. & Weissman, J. S. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* **484**, 538-541, (2012).

16 Elf, J., Nilsson, D., Tenson, T. & Ehrenberg, M. Selective charging of tRNA isoacceptors explains patterns of codon usage. *Science* **300**, 1718-1722, (2003).

17 Ran, W., Kristensen, D. M. & Koonin, E. V. Coupling between protein level selection and codon usage optimization in the evolution of bacteria and archaea. *MBio* **5**, e00956-00914, (2014).

18 Quax, T. E. *et al.* Differential translation tunes uneven production of operon-encoded proteins. *Cell Rep* **4**, 938-944, (2013).

19 Dana, A. & Tuller, T. The effect of tRNA levels on decoding times of mRNA codons. *Nucleic Acids Res* **42**, 9171-9181, (2014).

20 Zhang, F., Saha, S., Shabalina, S. A. & Kashina, A. Differential arginylation of actin isoforms is regulated by coding sequence-dependent degradation. *Science* **329**, 1534-1537, (2010).

21 Sharp, P. M. & Li, W. H. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15**, 1281-1295, (1987).

22 Chen, G. F. & Inouye, M. Suppression of the negative effect of minor arginine codons on gene expression; preferential usage of minor codons within the first 25 codons of the Escherichia coli genes. *Nucleic Acids Res* **18**, 1465-1473, (1990).

23 Zhang, G. *et al.* Global and local depletion of ternary complex limits translational elongation. *Nucleic Acids Res* **38**, 4778-4787, (2010).

24 Zhang, S. P., Zubay, G. & Goldman, E. Low-usage codons in Escherichia coli, yeast, fruit fly and primates. *Gene* **105**, 61-72, (1991).

25 Ikemura, T. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *J Mol Biol* **151**, 389-409, (1981).

26 Muramatsu, T. *et al.* Codon and amino-acid specificities of a transfer RNA are both converted by a single post-transcriptional modification. *Nature* **336**, 179-181, (1988).

27 Vivanco-Dominguez, S. *et al.* Protein synthesis factors (RF1, RF2, RF3, RRF, and tmRNA) and peptidyl-tRNA hydrolase rescue stalled ribosomes at sense codons. *J Mol Biol* **417**, 425-439, (2012).

28 Cruz-Vera, L. R., Magos-Castro, M. A., Zamora-Romo, E. & Guarneros, G. Ribosome stalling and peptidyl-tRNA drop-off during translational delay at AGA codons. *Nucleic Acids Res* **32**, 4462-4468, (2004).

29 Dong, H., Nilsson, L. & Kurland, C. G. Co-variation of tRNA Abundance and Codon Usage inEscherichia coliat Different Growth Rates. *Journal of Molecular Biology* **260**, 649-663, (1996).

30 Zamora-Romo, E., Cruz-Vera, L. R., Vivanco-Dominguez, S., Magos-Castro, M. A. & Guarneros, G. Efficient expression of gene variants that harbour AGA codons next to the initiation codon. *Nucleic Acids Res* **35**, 5966-5974, (2007).

31 Castillo-Mendez, M. A., Jacinto-Loeza, E., Olivares-Trejo, J. J., Guarneros-Pena, G. & Hernandez-Sanchez, J. Adenine-containing codons enhance protein synthesis by promoting mRNA binding to ribosomal 30S subunits provided that specific tRNAs are not exhausted. *Biochimie* **94**, 662-672, (2012).

32 Dittmar, K. A., Sorensen, M. A., Elf, J., Ehrenberg, M. & Pan, T. Selective charging of tRNA isoacceptors induced by amino-acid starvation. *EMBO Rep* **6**, 151-157, (2005).

33 Tuller, T. *et al.* An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141**, 344-354, (2010).

34 Bulmer, M. The selection-mutation-drift theory of synonymous codon usage. *Genetics* **129**, 897-907, (1991).

35 Akashi, H. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. *Genetics* **136**, 927-935, (1994).

36 Kramer, E. B. & Farabaugh, P. J. The frequency of translational misreading errors in E. coli is largely determined by tRNA competition. *RNA* **13**, 87-96, (2007).

37 Ninio, J. Fine tuning of ribosomal accuracy. *FEBS Lett* **196**, 1-4, (1986).

38 Drummond, D. A. & Wilke, C. O. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**, 341-352, (2008).

39 Acton, T. B. *et al.* Robotic cloning and Protein Production Platform of the Northeast Structural Genomics Consortium. *Methods Enzymol* **394**, 210-243, (2005).

40 Iost, I. & Dreyfus, M. The stability of Escherichia coli lacZ mRNA depends upon the simultaneity of its synthesis and translation. *Embo j* **14**, 3252-3261, (1995).

41 Iost, I., Guillerez, J. & Dreyfus, M. Bacteriophage T7 RNA polymerase travels far ahead of ribosomes in vivo. *J Bacteriol* **174**, 619-622, (1992).

42 Reuter, J. S. & Mathews, D. H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11**, 129, (2010).

43 Caliskan, N., Katunin, Vladimir I., Belardinelli, R., Peske, F. & Rodnina, Marina V. Programmed –1 Frameshifting by Kinetic Partitioning during Impeded Translocation. *Cell* **157**, 1619-1631, (2014).

44 Ieong, K. W., Pavlov, M. Y., Kwiatkowski, M., Forster, A. C. & Ehrenberg, M. Inefficient delivery but fast peptide bond formation of unnatural L-aminoacyl-tRNAs in translation. *J Am Chem Soc* **134**, 17955-17962, (2012).

45 Johansson, M., Zhang, J. & Ehrenberg, M. Genetic code translation displays a linear trade-off between efficiency and accuracy of tRNA selection. *Proc Natl Acad Sci U S A* **109**, 131-136, (2012).

46 Zaher, H. S. & Green, R. Quality control by the ribosome following peptide bond formation. *Nature* **457**, 161-166, (2009).

47 Chen, H., Shiroguchi, K., Ge, H. & Xie, X. S. Genome-wide study of mRNA degradation and transcript elongation in Escherichia coli. *Mol Syst Biol* **11**, 781, (2015).

48 Pelechano, V., Wei, W. & Steinmetz, Lars M. Widespread Co-translational RNA Decay Reveals Ribosome Dynamics. *Cell* **161**, 1400-1412, (2015).

49 Presnyak, V. *et al.* Codon optimality is a major determinant of mRNA stability. *Cell* **160**, 1111-1124, (2015).

50 dos Reis, M., Savva, R. & Wernisch, L. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res* **32**, 5036-5044, (2004).

51 Deana, A., Ehrlich, R. & Reiss, C. Synonymous codon selection controls in vivo turnover and amount of mRNA in Escherichia coli bla and ompA genes. *J Bacteriol* **178**, 2718-2720, (1996).

52 Nogueira, T., de Smit, M., Graffe, M. & Springer, M. The relationship between translational control and mRNA degradation for the Escherichia coli threonyl-tRNA synthetase gene. *J Mol Biol* **310**, 709-722, (2001).

53 dos Reis, M. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome. *Nucleic Acids Research* **31**, 6976-6985, (2003).

54 Ude, S. *et al.* Translation elongation factor EF-P alleviates ribosome stalling at polyproline stretches. *Science* **339**, 82-85, (2013).

55 Doerfel, L. K. *et al.* EF-P is essential for rapid synthesis of proteins containing consecutive proline residues. *Science* **339**, 85-88, (2013).

56 Ivanova, N., Pavlov, M. Y. & Ehrenberg, M. tmRNA-induced release of messenger RNA from stalled ribosomes. *J Mol Biol* **350**, 897-905, (2005).

57 Richards, J., Sundermeier, T., Svetlanov, A. & Karzai, A. W. Quality control of bacterial mRNA decoding and decay. *Biochim Biophys Acta* **1779**, 574-582, (2008).

58 Shoemaker, C. J., Eyler, D. E. & Green, R. Dom34:Hbs1 promotes subunit dissociation and peptidyl-tRNA drop-off to initiate no-go decay. *Science* **330**, 369-372, (2010).

59 Becker, T. *et al.* Structural basis of highly conserved ribosome recycling in eukaryotes and archaea. *Nature* **482**, 501-506, (2012).

60 Chadani, Y., Ono, K., Kutsukake, K. & Abo, T. Escherichia coli YaeJ protein mediates a novel ribosome-rescue pathway distinct from SsrA- and ArfA-mediated pathways. *Mol Microbiol* **80**, 772-785, (2011).

61 Chevrier-Miller, M., Jacques, N., Raibaud, O. & Dreyfus, M. Transcription of single-copy hybrid lacZ genes by T7 RNA polymerase in Escherichia coli: mRNA synthesis and degradation can be uncoupled from translation. *Nucleic Acids Res* **18**, 5787-5792, (1990).

62 Leroy, A., Vanzo, N. F., Sousa, S., Dreyfus, M. & Carpousis, A. J. Function in Escherichia coli of the non-catalytic part of RNase E: role in the degradation of ribosome-free mRNA. *Molecular Microbiology* **45**, 1231-1243, (2002).

63 Marchand, I., Nicholson, A. W. & Dreyfus, M. Bacteriophage T7 protein kinase phosphorylates RNase E and stabilizes mRNAs synthesized by T7 RNA polymerase. *Mol Microbiol* **42**, 767-776, (2001).

64 Cardinale, C. J. *et al.* Termination factor Rho and its cofactors NusA and NusG silence foreign DNA in E. coli. *Science* **320**, 935-938, (2008).

65 Proshkin, S., Rahmouni, A. R., Mironov, A. & Nudler, E. Cooperation between translating ribosomes and RNA polymerase in transcription elongation. *Science* **328**, 504-508, (2010).

66 Kimchi-Sarfaty, C. *et al.* A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science* **315**, 525-528, (2007).

67 Li, X., Hirano, R., Tagami, H. & Aiba, H. Protein tagging at rare codons is caused by tmRNA action at the 3' end of nonstop mRNA generated in response to ribosome stalling. *RNA* **12**, 248-255, (2006).

68 Plotkin, J. B. & Kudla, G. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet* **12**, 32-42, (2011).

69 Chiba, S. & Ito, K. Multisite ribosomal stalling: a unique mode of regulatory nascent chain action revealed for MifM. *Mol Cell* **47**, 863-872, (2012).

70 Letzring, D. P., Wolf, A. S., Brule, C. E. & Grayhack, E. J. Translation of CGA codon repeats in yeast involves quality control components and ribosomal protein L1. *RNA* **19**, 1208-1217, (2013).

71 Ramu, H. *et al.* Nascent peptide in the ribosome exit tunnel affects functional properties of the A-site of the peptidyl transferase center. *Mol Cell* **41**, 321-330, (2011).

72 Sorensen, M. A. *et al.* Over expression of a tRNA(Leu) isoacceptor changes charging pattern of leucine tRNAs and reveals new codon reading. *J Mol Biol* **354**, 16-24, (2005).

73 Gao, W., Tyagi, S., Kramer, F. R. & Goldman, E. Messenger RNA release from ribosomes during 5'-translational blockage by consecutive low-usage arginine but not leucine codons in Escherichia coli. *Mol Microbiol* **25**, 707-716, (1997).

74 Ito, K. & Chiba, S. Arrest peptides: cis-acting modulators of translation. *Annu Rev Biochem* **82**, 171-202, (2013).

75 Tsai, Y. C. *et al.* Recognition of the 70S ribosome and polysome by the RNA degradosome in Escherichia coli. *Nucleic Acids Res* **40**, 10417-10431, (2012).

76 Dreyfus, M. Chapter 11 Killer and Protective Ribosomes. **85**, 423-466, (2009).